Published online 22 May 2009 Nucleic Acids Research, 2009, Vol. 37, Web Server issue W317–W322
doi:10.1093/nar/gkp416

GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information

Ruben Nogales-Cadenas[1], Pedro Carmona-Saez[1], Miguel Vazquez[2], Cesar Vicente[1], Xiaoyuan Yang[1], Francisco Tirado[1], Jose María Carazo[3] and Alberto Pascual-Montano[1,*]

[1]Computer Architecture Department, [2]Software Engineering Department, Complutense University of Madrid and [3]Biocomputing Unit, National Center for Biotechnology, CNB-CSIC, Madrid, Spain

Received January 31, 2009; Revised April 24, 2009; Accepted May 6, 2009

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 24, 2015

ABSTRACT

GeneCodis is a web server application for functional analysis of gene lists that integrates different sources of information and finds modular patterns of interrelated annotations. This integrative approach has proved to be useful for the interpretation of high-throughput experiments and therefore a new version of the system has been developed to expand its functionality and scope. GeneCodis now expands the functional information with regulatory patterns and user-defined annotations, offering the possibility of integrating all sources of information in the same analysis. Traditional singular enrichment is now permitted and more organisms and gene identifiers have been added to the database. The application has been re-engineered to improve performance, accessibility and scalability. In addition, GeneCodis can now be accessed through a public SOAP web services interface, enabling users to perform analysis from their own scripts and workflows. The application is freely available at http://genecodis.dacya.ucm.es

INTRODUCTION

High-throughput experiments such as DNA microarrays or proteomics techniques have been widely used during the last decade. Currently they are standard technologies in many research centers and facilities. Although these methodologies generate huge amounts of information, the challenge lies not only in the data processing area, in which the bioinformatics community has made significant advances, but also in the interpretation of the information. In this context, new data mining techniques able to extract interpretable facts and biological knowledge are needed to understand the biological meaning of experiments.

An essential task in this analysis is to translate gene signatures into information that can assist in understanding the underlying biological mechanisms. In the last few years several methods and tools have been developed to interpret large lists of genes or proteins using information available in biological databases. The common idea used in most of these methods is to find functional descriptors that are significantly enriched in the gene signature.

The first type of techniques that emerged in this field were focused on evaluating the frequency of individual annotations and apply an statistical test to determine which annotations are significantly enriched in an input list with respect to a reference list, usually the whole genome or all genes in a microarray. Annotations from different sources like Gene Ontology (GO) (1) or KEGG (2) are commonly used in this context. Several tools have been developed following this approach, and although each application introduces its own variations such as different statistical tests, sources of annotations or supported organisms, they all carry out the same type of analysis, producing only slightly differences in the results. Good reviews of such methods can be found in (3,4).

A fresh line of research appeared with the observation that the use of thresholds to select the significant genes could underestimate the effect of significant biological effects during the functional analysis. This idea derived in a new and different analytical concept in which the distribution of annotations is evaluated over the whole list of genes, sorted by their correlation with the phenotype. In this context, the gene set enrichment analysis (GSEA) was introduced by Subramanian and Tamayo (5) and since then different methods have followed this approach.

Nevertheless both approaches, the standard enrichment analysis and GSEA methods evaluate each annotation

*To whom correspondence should be addressed. Tel: +34 913944420; Fax: +34 913944687; Email: pascual@fis.ucm.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

independently from the others without taking into account the potential relationships among them. However, most of the annotations in biological databases are interconnected because they are associated to common genes. Patterns that contain these relationships can provide invaluable information and extend our understanding of biological events associated to the experimental system. It is therefore highly desirable to evaluate these associations in the functional analysis of gene lists. New tools that attempt to extract this type of information have been proposed [see a review in (6)].

In 2007, we introduced GeneCodis (7), a tool for modular enrichment analysis oriented to integrate information from different sources and find enriched combinations of annotations in large lists of genes or proteins. Modular enrichment represents a step forward in the functional analysis because of its capacity to integrate heterogeneous annotations and to discover significant combinations among them. Since its original publication, this tool has achieved more than 25 000 submissions from all over the world and has been referenced in different works. A mirror has even been recently set up at the Center for Bioinformatics in Peking University to facilitate the access in the Asian region.

In this work, we present a new version of the software with improved functionality, performance, accessibility and extended scope. First, we have expanded the type of information available in the application by incorporating new types of annotations such as microRNAs and transcription factors. In this way, the new version of GeneCodis offers now the possibility to mine not only functional information but also regulatory patterns with the potential to integrate both sources of annotations in the same analysis. Moreover, the application allows researchers to submit their own annotations and perform a joint analysis with the rest of available information. The new version also includes the analysis of individual annotations (singular enrichment analysis) and new type of gene identifiers and organisms were included to cover more possible analytical scenarios.

From a technical point of view, GeneCodis has been completely reengineered making it faster and more flexible. The algorithm that retrieves sets of concurrent annotations has been improved and runs in a multi-grid environment to better handle large number of jobs simultaneously and thus improving the performance and the throughput of the system. The application can now be accessed in a programmatic way using SOAP Web Services. This allows researchers to include GeneCodis functionalities in their own data analysis pipelines. Finally, the new friendly interface design facilitates the use of the tool, and new graphs and file formats have been added in the results to facilitate its processing and interpretation.

FEATURES AND FUNCTIONALITY

GeneCodis has suffered plenty of changes since its first publication, but the tool still works in a similar way. Users have to select the organism and the biological annotations to be considered in the analysis and then load the gene list. As advanced options the reference set of genes can be specified; otherwise the whole genome will be used by default. It is also possible to select the preferred statistical test among three possible alternatives: hypergeometric test (default), chi-square test or both. Annotations will be considered in the concurrent analysis only if they appear in at least a minimum number of genes, known as minimum support, which is set to three by default. Finally, users can select the multiple hypothesis correction method: false discovery rate method (default), permutation-based correction or none. As new feature, we allow the submission of a file with a list of user-defined annotations that can be considered in the analysis together with the rest of selected annotations.

Functional analysis in GeneCodis

As stated previously, the enrichment analysis of individual annotations was the first method introduced for the functional interpretation of large lists of genes or proteins, and still remains the most popular one. There is a large collection of tools that implement this methodology, most of them focused on the evaluation of GO annotations. The arrival of GSEA methods turned the enrichment analysis of individual annotations from a gene-centered to a gene-set based analysis. These methods, although extremely useful for the interpretation of gene lists, do not exploit the inter-relationships that exist among gene annotations. In this context, tools such as GeneCodis provide a new way to analyze functional information by taking into account these relationships among annotations associated to common genes in the list. This analysis offers different advantages with respect to singular enrichment methods. Combined terms may contain unique biological meaning for a given study, not held by individual terms. For example, the combination of two terms such as *Apoptosis* and *Mitochondria* may be enriched in a gene list while the individual annotations are not significant if evaluated independently in the same list of genes. But probably more interesting is the potential of this approach to integrate and jointly analyze information that covers different aspects of the biology of the genes. Although there are several applications for modular analysis [see (6) for a review], GeneCodis is one of the few tools that tackle this problem by offering enrichment analysis of concurrent annotations using an algorithm based on the extraction of frequent itemsets [see (8) for details].

In this new version of GeneCodis, we have also included the singular enrichment. That is, in addition to the analysis of combinations of annotations, analysis of individual annotations is also performed and included in the results. In this way, if two categories are selected, results will include three different lists: one with the concurrent analysis results and two other lists with singular enrichment information, one for each category. Results are provided in different formats: html tables, tabulated text files, xml files, pie charts and bar graphs.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 24, 2015

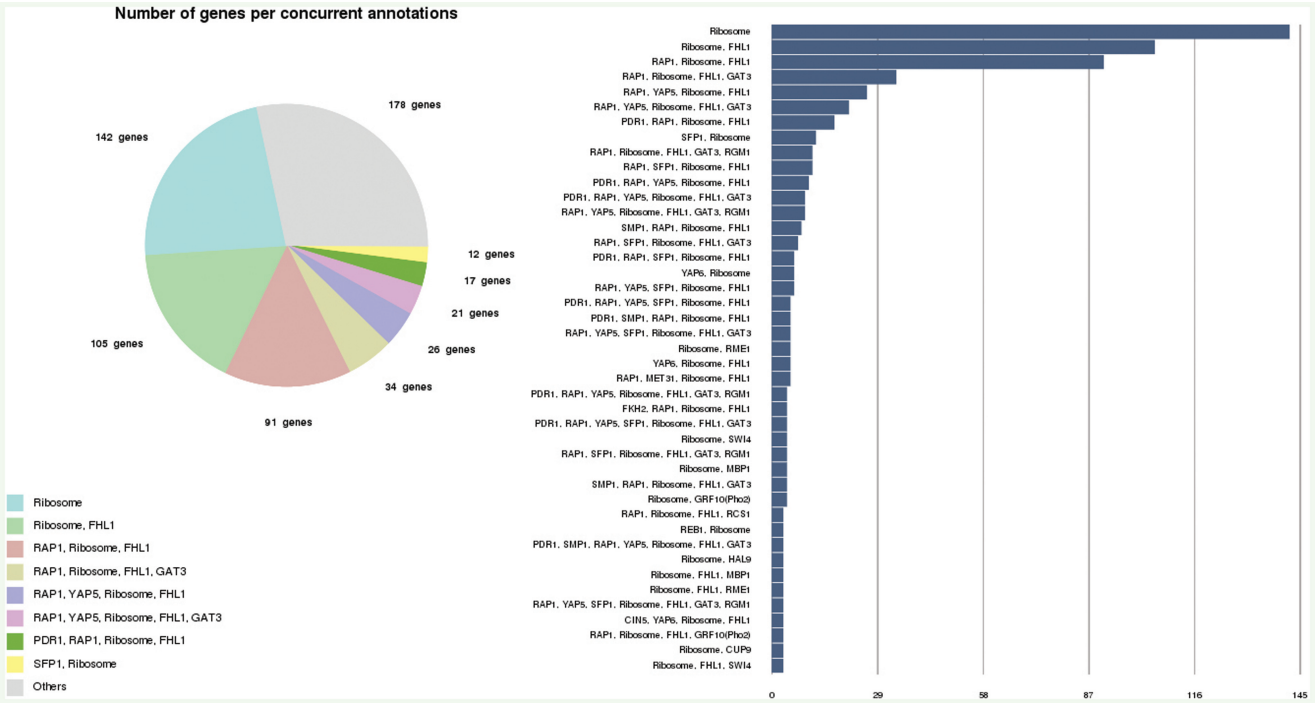

Figure 1. Example of the graphics generated in a typical GeneCodis analysis. Enriched combinations of significant annotations are represented in a pie and bar graphs, where the length of the bars and size of the slices are proportional to the number of genes supporting the significant combination of annotations.

New sources of annotations: integrating functional and regulatory information

Among the multiple resources of information, GO is by far the most popular one for functional analysis of gene lists. This is reasonable due to the rich content of GO in terms that describe the functional role of genes at a molecular level, and the initiatives of different consortiums to annotate complete genomes with GO terms. Earlier enrichment tools were mainly based on the analysis of GO terms, although information from other sources such as KEGG or Biocarta (www.biocarta.org) has been incorporated in more recent applications. Even though GO covers three aspects of the biology of genes: Biological Process, Cellular Component and Molecular Function, the former is the information in which researchers have been mainly interested for the functional characterization of gene lists. This is in part because biological process terms provide explicit information to interpret the biological mechanisms that may be associated to the experimental system.

Nevertheless, beyond functional information there are other properties of genes and proteins that can also be very useful to interpret biological systems, such as information related to gene expression regulation. There are different sources of regulatory information that have been incorporated in enrichment tools (9–11). In GeneCodis, we now include annotations related to regulatory information from different places. For example, miRBase (12) that contains putative targets of microRNAs, molecules that in the last few years have been shown as key regulators in many biological systems. From this database we have extracted the microRNAs associated to genes in different organisms: *B. taurus, C. elegans, D. rerio, D. melanogaster, G. gallus, H. sapiens, M. musculus* and *R. norvegicus*. Another source is transcription factors, which are key regulators that provide significant information in enrichment analysis [e.g. (5,13,14)]. Like in the case of GSEA (5) or FactorY (13), we have annotated genes from *H. sapiens, M. musculus* and *R. norvegicus* based on the information contained in TransFac database. In addition, we have used results from chip-on-chip experiments (15) to annotate genes in *S. cerevisiae* with transcription factors that bind to their promoter regions. These new annotations allow users to perform enrichment analysis with regulatory information in addition to functional information.

Besides the independent analysis of different properties of genes, the integration of heterogeneous sources of information can provide a more complete picture of the system under analysis. In this context, the new sources of information offer the possibility to mine gene lists to discover associations among regulatory and functional information. This integration is probably one of its most useful features. This is illustrated in Figure 1 with a very simple example. It shows a screenshot of the GeneCodis results from the analysis of the ribosomal gene set in KEGG. Using the information of transcription factors in the analysis we can see how co-annotations of functional and regulatory information are evaluated. For example enriched annotations contain transcription factors such as RAP1, FHL1 or SFP1 that are well known to play important roles in the regulation of ribosomal genes.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 24, 2015

Organisms and identifiers

To extend the scope of GeneCodis we have also increased the number of supported organisms and gene identifiers. GeneCodis works with most of the model organisms in biological research. The whole list includes *Arabidopsis thaliana, Bos taurus, Caenorhabditis elegans, Candida albicans, Danio rerio, Drosophila melanogaster, Escherichia coli, Gallus gallus, Homo sapiens, Leishmania major, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae, Schizosaccharomyces pombe* and *Vibrio cholerae*. Alternatively, to facilitate the usability of the application we have extended the type of gene identifiers that are supported, including propietary identifiers from commercial platforms such as Affymetrix, Agilent, Codelink and Illumina. The backbone of the system is now based on Ensembl and gene ids have been obtained using biomart (http://www.biomart.org). A new menu is now available for the selection of the reference list that includes arrays from the most popular manufacturers: Affymetrix, Agilent and Illumina.

Interface

The new friendly interface has been designed to facilitate usability. Results can be explored more dynamically because they are presented in different formats, including pie charts and bar graphs images that allow users to interpret the data with a quick look (see Figure 1 for example). XML files with structured results are also provided permitting its use in data processing pipelines. Tabulated text files and html tables are also available. These different file formats allow researchers to use GeneCodis results in other applications.

Big efforts have been done to facilitate the access of GeneCodis functionality by different ways. In addition to the classical web-server access, the tool can now be used programmatically trough the SOAP Web Services technology. Using this technology, researchers can insert functional analysis in their data mining pipelines or in other bioinformatics systems in a very straightforward manner (see Figure 2 for a very simple code example). There are many advantages in the use of Web Services. It is a platform-independent and language-independent technology which makes it very adequate for loosely coupled systems working in different architectures. A complete tutorial can be found in the web site including example scripts in Ruby, Perl and PHP.

IMPLEMENTATION

The algorithmic core of GeneCodis has been reviewed and drastically improved. There is a new implementation of the method to extract closed itemsets based on a modification of previously reported efficient techniques (16). The new algorithm can deal with large sets of annotations in a faster and more efficient way than the methodology implemented in the previous version. This is especially evident if multiple annotations are included in the same analysis or when the permutation-based test is used for the multiple hypothesis correction. In both cases the computing time has decreased drastically with respect to the previous version.

The throughput and performance of the entire system has also been improved by implementing the new algorithm and the Web Service technology in the context of a multi-grid computational environment. The new system is able to handle all submitted jobs simultaneously without queuing any of them for long periods of time. The current implementation takes advantage of one cluster with two Quad-Core Intel Xeon processors of 64 bits and two independent grid infrastructures (CyTED, http://www.cyted.org and EELA2, http://www.eu-eela.org/) integrated by the GridWay metascheduler (17).

When users submit a query through the web site, the system launches one job for the concurrent analysis and one job for each selected annotation category in the singular enrichment. All jobs will be executed in parallel. Users can also submit a query directly from a script using the Web Services. In all cases the workflow of the system is the same: a meta-scheduler determines, depending on the computational load of the cluster and the grids, in which computational environment the jobs will be executed. The jobs are queued and executed in the cluster if it is free. Otherwise, the jobs are sent to the less work-loaded grid resource. This approach represents a cost-effective alternative to improve the throughput of the application, and to guarantee its real-time performance, a critical aspect of any popular web-server application.

CONCLUSIONS AND DISCUSSION

High-throughput experimental techniques have demonstrated to be very useful for the study of biological systems from a global perspective. In many cases, these techniques generate huge amounts of data in the form of large gene or protein lists. An essential task in this context is to translate these lists to functional information that aids researchers in the interpretation of the underlying biological processes. Although this interpretation is not always trivial, methods based on the enrichment analysis have been successfully used for this purpose.

GeneCodis is a tool designed to expand the traditional enrichment analysis of annotations by adding the possibility of evaluating not only individual terms, but also their significant combinations. Since its creation, this application has become a useful resource for the research activity. This encourages us to improve it by adding more functionality to extend its scope, performance and accessibility.

In summary, the new version of GeneCodis finds combinations of annotations and also includes the traditional singular enrichment analysis. New annotations are now available, such as microRNAs or transcription factors which help to expand the functionality of GeneCodis towards the analysis of regulatory information. Moreover, GeneCodis allows researchers to jointly analyze regulatory and functional information and to extract association patterns between both data sources. The algorithm that retrieve sets of concurrent annotations has been improved and implemented for running in a multi-grid

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 24, 2015

```ruby
require 'soap/wsdlDriver'

#### Test arguments ###
#
org = "Sc" #Saccharomyces Cerevisiae
algorithm = 1 # Concurrent analysis
test = 0 # Hypergeometric p values
correction = 1 # FDR method
minsupport = 3 # Minimum support of 3 is considered in the analysis
annotations = ["GO_Biological_Process","GO_Molecular_Function"]
reference_list = [] # whole genome
input_list = ["S000004295", "S000005284", "S000000598", "S000006205",
              "S000006183", "S000004982", "S000005662", "S000001387",
              "S000002555", "S000005668", "S000002585", "S000003736"] #input list


#### Connecting to server ###
#
WSDL_URL = "http://genecodis.dacya.ucm.es/static/wsdl/genecodisWS.wsdl"
driver = SOAP::WSDLDriverFactory.new(WSDL_URL).create_rpc_driver

### Submit tha analysis ###
#
job_id = driver.analyze(org,algorithm,test,correction,minsupport,input_list,annotations,reference_list)

### Check job state ###
#
status = driver.status(job_id)
while status == 1
  puts "Waiting ..."
  sleep 2
  status = driver.status(job_id)
end

# If error
if status < 0
  error_message = driver.info(job_id)
  raise "Finished with error #{ error_message }"
end

# Get results
results = driver.results(job_id)
puts results
```

Figure 2. Example of a ruby client code to invoke the GeneCodis Web Service. The access only needs three main steps: submit the analysis, ask for the job status and get the results.

environment that is able to handle all submitted jobs simultaneously. In this way the performance and the throughput of the system is enhanced. Finally, GeneCodis can be accessed programmatically by a web services interface; allowing its inclusion in data analysis pipelines.

One of the disadvantages of the modular enrichment analysis included in GeneCodis is the intrinsic redundancy of the combined functional annotations. This is caused by the unavoidable natural redundancy of the information across biological databases and by the nature of the data mining algorithm that extract all possible combinations. A method to filter out and group annotations that are related to the same biological topic is still needed. Going in the direction of creating a more self-contained application, new future releases will also include GSEA methods to allow users the selection of all possible flavours of functional analysis in the same environment.

The new version has been running since August 2008. Extensive tests have been carried out using synthetic and real datasets for which the outcome of the software is known. The diverse functionalities supported by this tool have been also fully tested by real users who have provided feedback on issues that have helped in improving the application. We hope the renewed GeneCodis will be of interest to the scientific community.

GENECODIS AVAILABILITY

This application can be freely accessed through its main site at http://genecodis.dacya.ucm.es. A mirror has also been recently set up at the Center for Bioinformatics

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 24, 2015

in Peking University. The mirror is available at http://genecodis.cbi.pku.edu.cn/.

ACKNOWLEDGEMENTS

The authors thank the support of Integromics, S.L. This work also makes use of results produced by the EELA-2 project (www.eu-eela.eu), co-funded by the European Commission within its Seventh Framework Programme. The authors like to acknowledge Luis Canet for his technical help. Special thanks also to Prof. Jinchu Luo from the Center for Bioinformatics at Peking University for his help in setting the Asian GeneCodis mirror. A.P.M. acknowledges the support of the Spanish Ramón y Cajal program.

FUNDING

Spanish grants BIO2007-67150-C03-02, S-Gen-0166/2006, CYTED-505PI0058, TIN2005-5619 and PS-010000-2008-1 and European Union Grant FP7-HEALTH-F4-2008-202047. Funding for open access charge: Spanish Grant BIO2007-67150-C03-02.

Conflict of interest statement. None declared.

REFERENCES

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
2. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
3. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics,*, **21**, 3587–3595.
4. Dopazo,J. (2006) Functional interpretation of microarray experiments. *OMICS*, **10**, 398–410.
5. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
6. Huang da,W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
7. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.
8. Carmona-Saez,P., Chagoyen,M., Rodriguez,A., Trelles,O., Carazo,J.M. and Pascual-Montano,A. (2006) Integrated analysis of gene expression by Association Rules Discovery. *BMC Bioinformatics*, **7**, 54.
9. Abascal,F., Carmona-Saez,P., Carazo,J.M. and Pascual-Montano,A. (2008) ChIPCodis: mining complex regulatory systems in yeast by concurrent enrichment analysis of chip-on-chip data. *Bioinformatics,*, **24**, 1208–1209.
10. Dennis,G. Jr, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.
11. Al-Shahrour,F., Minguez,P., Tarraga,J., Montaner,D., Alloza,E., Vaquerizas,J.M., Conde,L., Blaschke,C., Vera,J. and Dopazo,J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, **34**, W472–W476.
12. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
13. Guruceaga,E., Segura,V., Corrales,F.J. and Rubio,A. (2009) FactorY, a bioinformatic resource for genome-wide promoter analysis. *Comput. Biol. Med.*, **39**, 385–387.
14. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
15. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, **298**, 799–804.
16. Zaki,M.J. and Hsiao,C.-J. (2005) Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure. *IEEE Trans. Knowledge Data Eng.*, **17**, 12.
17. Huedo,E., Montero,R.S. and Llorente,I.M. (2005) The GridWay framework for adaptive scheduling and execution on grids. *SCPE*, **6**, 8.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 24, 2015